Human knowledge depends not only on textual information represented by word symbols that do not inherently have semantic meaning, but also on perceptual information, such as visual and acoustic cues. I believe incorporating multiple modalities is an important step towards general artificial intelligence. My research interests are in natural language processing (NLP), particularly in its fusion with computer vision, and its interpretability and generalizability. I have two main research directions:

- **NLP & Multimodal Learning**: How can we combine language with other modalities better?
- **Interpretability & Generalizability**: How can we improve the interpretability and generalizability of NLP systems and multimodal systems?

I have been working with Professors Ellie Pavlick and Chen Sun from the inception of my graduate program at Brown University, and I have been fortunate to work on research projects relevant to these directions.

NLP & Multimodal Learning Even though large-scale pretrained language models have shown strong empirical performance on natural language understanding benchmarks, they have been criticized for their lack of grounding, or the connections between words and real-world entities (Bender and Koller, 2020). I would like to address this criticism by jointly training models on language data and data of other modalities, which enables the surface form of language to ground to entities in real world.

One of the multimodal models is vision-and-language (VL) models. Prior evaluations of VL models focus mostly on inherently multimodal tasks, such as visual question answering, but it remains unclear whether VL pretraining yields better linguistic representations. This question is investigated in our published work at Findings of EMNLP 2021, where I am the first author (Yun et al., 2021). We carefully pretrain VL models and their text-only variants in a controlled manner, and compare the semantic representations learned via VL and text-only pretraining using a suite of analyses. In all, our findings indicate that VL pretraining sometimes produces gains over text-only pretraining, but the margins are too insignificant to support the conclusion that VL pretraining in its current form can bring benefits to NLP in general.

Nonetheless, it is possible that the current physical commonsense reasoning evaluation datasets require high-level compositional understanding of different objects, adjectives, and verbs, whereas VL models might have learned low-level isolated concepts, such as shapes and spatial directions. Thus, I recently proposed a research project at BigScience Workshop organized by Hugging Face. This project aims to build a benchmark to evaluate how well language models understand the word relations and word meanings. For example, can language models distinguish the word forms "red" and "blue", or can they link the word forms to their meanings in real world? We are currently constructing tests for low-level concepts that are represented in world representations. This benchmark would not only be a grounding-ness benchmark for language models, but also a meaningful one to evaluate the linguistic representations learned by multimodal pretraining which involves language.

There are two other directions that I am interested in. First, how do multimodal systems perform in interactive environments? I am currently working with Professor Stefanie Tellex on a situation-aware coaching dialogue system (CoachDial) that guides users to accomplish daily tasks by asking clarification questions. Second, whereas most of the VL models need to learn a classification layer for each downstream task, I would like to train an autoregressive VL model with multitask prompted training (Sanh, Webson, Raffel, and Bach et al., 2021) in order to

formulate downstream tasks as causal language modeling task and hopefully enable the model to perform tasks in a zero-shot fashion.

Interpretability & Generalizability End-to-end training is one of the dominant approaches to train neural networks, but it is challenging to understand their internal mechanism. I believe it is important to explore and improve the interpretability of these models, since such analyses provide insights into limitations and potentials of existing architectures. Also, interpretable models might allow transfer learning with few-shot or even zero-shot setting and yield better generalization (Koh et al., 2020; Mao et al., 2019).

Humans can compose learned primitive concepts (e.g., "purple" and "apple") into novel composite concepts (e.g., "purple apple"). In our recently work submitted to CVPR 2022, where I am the first author (Yun et al., 2021), we propose a framework for measuring how well pretrained VL models can learn composable primitive concepts and generalize to new composite concepts. Specifically, we focus on pretrained CLIP (Radford et al., 2021) and use its predicted primitives to train a derivation model that maps primitives to composite concepts. The experimental results reveal that the primitive concepts from CLIP are useful for visual recognition tasks, but the learned derivations are often not interpretable, indicating more research is needed to improve models' ability to capture primitive concepts. Deriving inspiration from the classical model of concepts in linguistics, I am interested in extending this research by designing a composition function that composes primitives to composite concepts using contrastive learning.

Conclusion I am strongly motivated to pursue a Ph.D., since it would allow me to do more focused research. I would like to keep evaluating existing multimodal models, and explore how to sufficiently combine complementary information from different modalities based on the observed limitations of VL models from my previous research. I plan to continue my research after my Ph.D., either as a faculty in academia or as a research scientist in industry.

At Brown, there are several professors whose projects are especially appealing to me. **Professor Ellie Pavlick**'s research about grounded language learning and understanding of internal mechanism of NLP systems interests me. Furthermore, I would like to be co-advised by an advisor with expertise in computer vision, and I am interested in working with **Professor Chen Sun** on multimodal representation learning and visual concept learning. I would also like to work with **Professor Stefanie Tellex** on incorporating language into interactive environments. After studying recent works of these professors, I can see a clear match of my skills and research interests at Brown, and I believe I will make productive and impactful contributions both to the Brown community, and the field at large.

Tian Yun, Chen Sun, Ellie Pavlick. Does Vision-and-Language Pretraining Improve Lexical Grounding? *Findings of EMNLP 2021*.

Tian Yun, Usha Bhalla, Ellie Pavlick, Chen Sun. Do Vision-Language Pretrained Models Learn Primitive Concepts? **Submitted to CVPR 2022.**

Emily Bender, Alexander Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *ACL 2020*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv preprint arXiv:2110.08207, 2021*.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. *ICLR 2019*.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, Percy Liang. Concept Bottleneck Models. *PMLR 2020*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Image, 2:T2, 2021*.